A Corpus Linguistics Primer

Through AntConc

What is a corpus?

What?

It's so simple.

It's just a lot of words!

Well, not exactly.

What is a corpus?

A machine readable collection of texts from spoken or written sources that were created in a natural expressive context.

Machine readable: text formats that can be loaded, parsed and manipulated independent of platforms. Despite being technically ‘open’, they can have dense annotations identifying various kinds of descriptive features. Typically these files are stored in the form of plain text files stored in ASCII or UTF encoding or structured XML files.

Natural expressive context: texts that were not created for the purpose of corpus analysis; in other words, texts that partake in an authentic communication.

A **machine readable** collection of texts from spoken or written sources that were created in a **natural expressive context.**

It constitutes methodology for studying the nature of language.

In such a collection, it is expected that there is an intention:

- To identify the collection as representative and balanced in the context of a language, variety, register, or genre. It has a purpose, in other words, so it should aim to reference what is typical.
- To analyse the collection linguistically (attention to word frequency, language change, morphemes, and the like), with explicit annotations.

Types of corpora

General: represents a language in a holistic way.

Specific: restricted to a particular variety, register, or purpose.

Raw: contains files of only corpus material (plain text)

Annotated: contains additional descriptive information (usually with metadata), encoded with parts-of-speech tags, or XML tags under the guidelines of the Text Encoding Initiative (TEI) or Corpus Encoding Standard (CES). Annotated corpuses include information *about* the text within *markup*. This kind of corpus can also be *lemmatized* (each word is followed by its lemma—–the standard form that you would look up in the dictionary).

More types of corpora

Diachronic: shows language change over time.

Synchronic: shows a snapshot of language in a time.

Monolingual: shows one language.

Parallel: shows the same text in multiple languages.

Static: have a fixed size (e.g. British National Corpus).

Dynamic: can be constantly extended (e.g. Bank of England).

Does this analysis constitute a theory of language?

Strictly speaking, no. These analyses offer loads of information about huge amounts of textual data, but they only offer information about frequencies. There is no straightforward semantic meaning in a corpora; what you are seeing in corpus analysis is:

- Frequencies of items (how often words or morphemes or grammatical structures occur in a text)
- Frequencies of co-occurring things (that is, groups of words or grammatical structures)

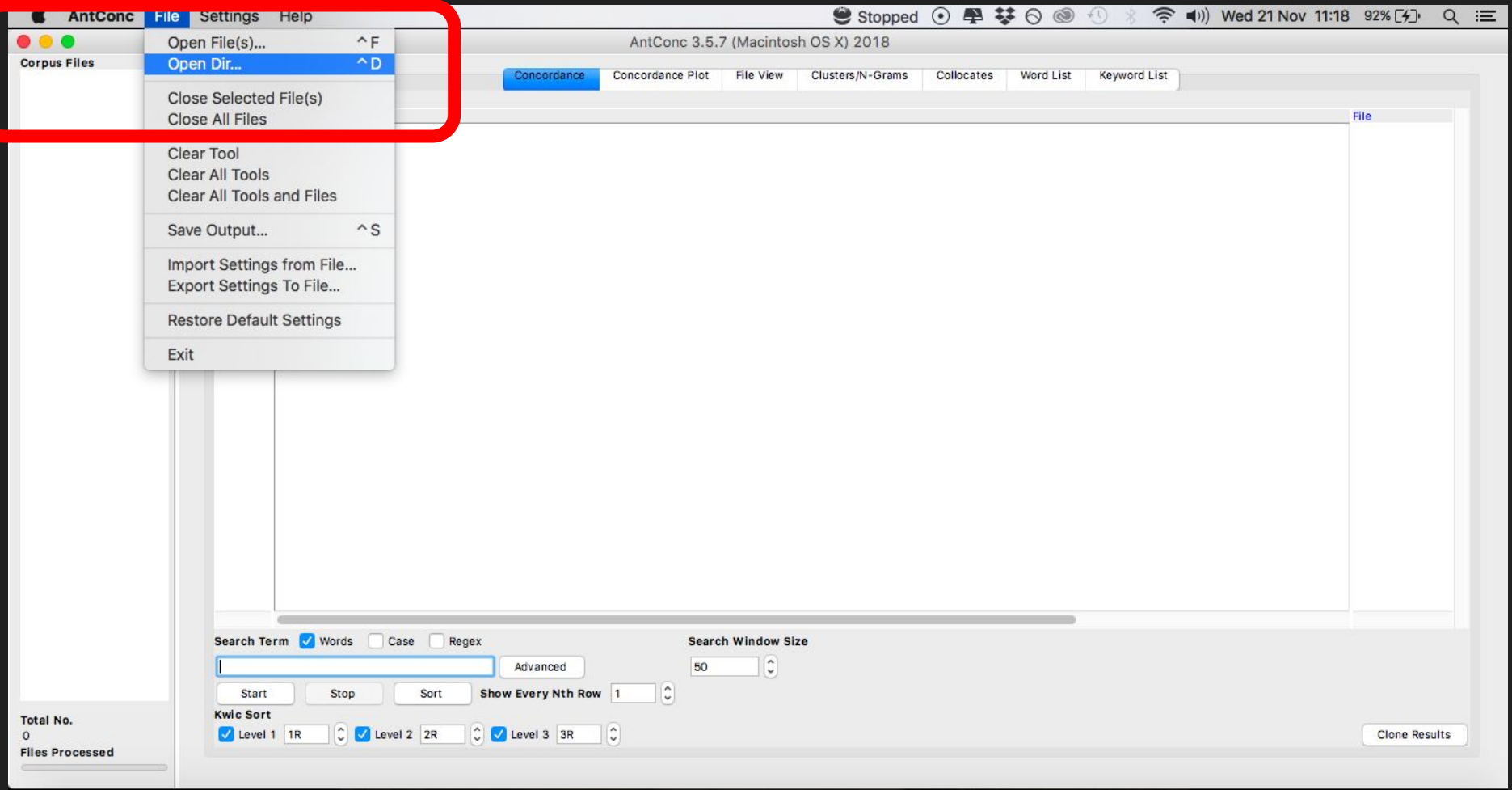
The work of interpretation still needs to be done; you need to decide what is meaningful. Remember, though, that ***formal differences reflect functional differences.*** Formal qualities illustrate functional regularities in communication.

AntConc <http://www.laurenceanthony.net/software.html>

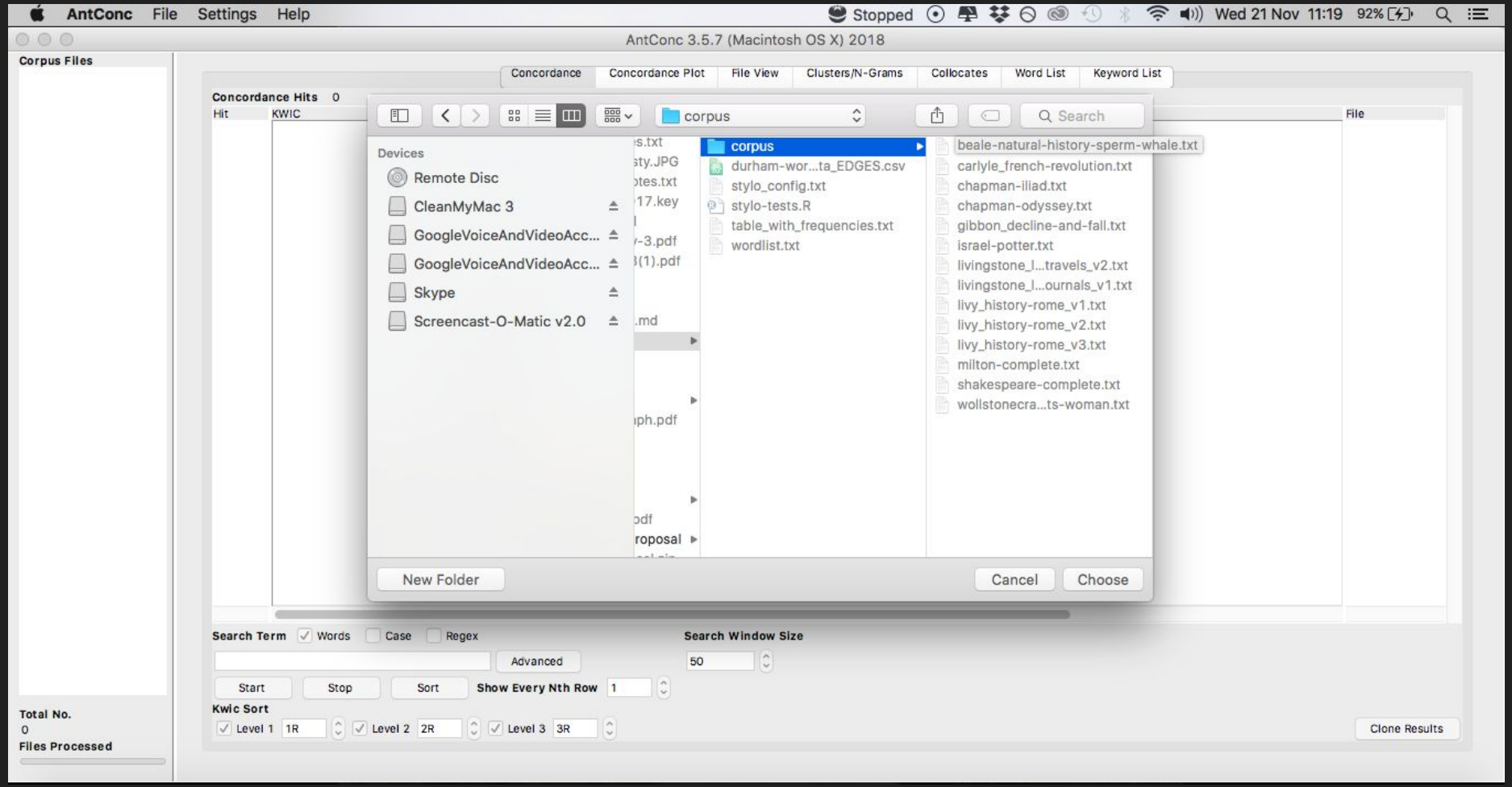
Pros: Free, well-maintained application; can search and analyse multiple texts in a corpus; has impressive key-word-in-context functions.

Cons: It can only perform basic corpus analyses (i.e., it cannot do more complex linguistic analyses).

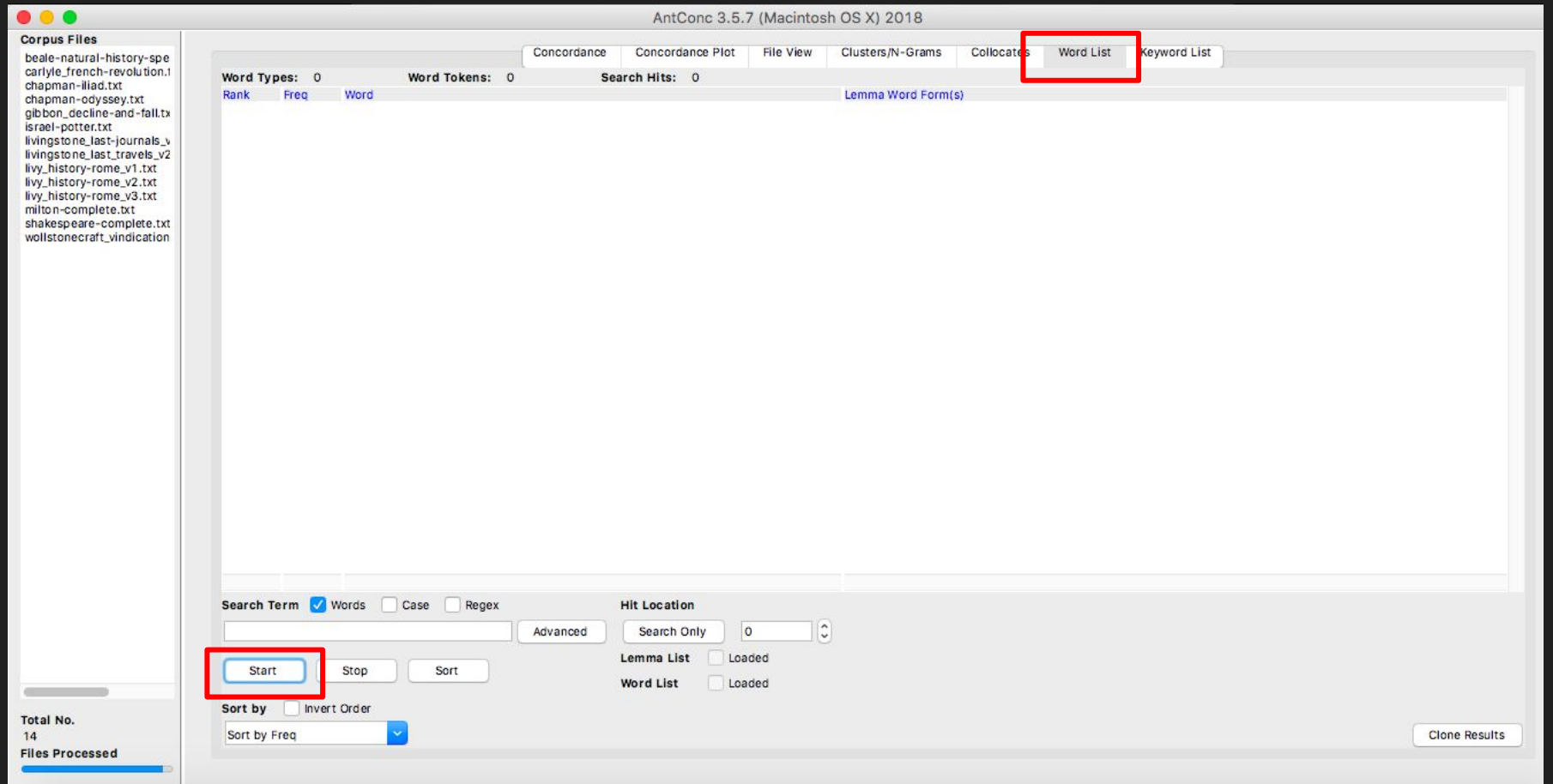
Getting started: File > OpenDir



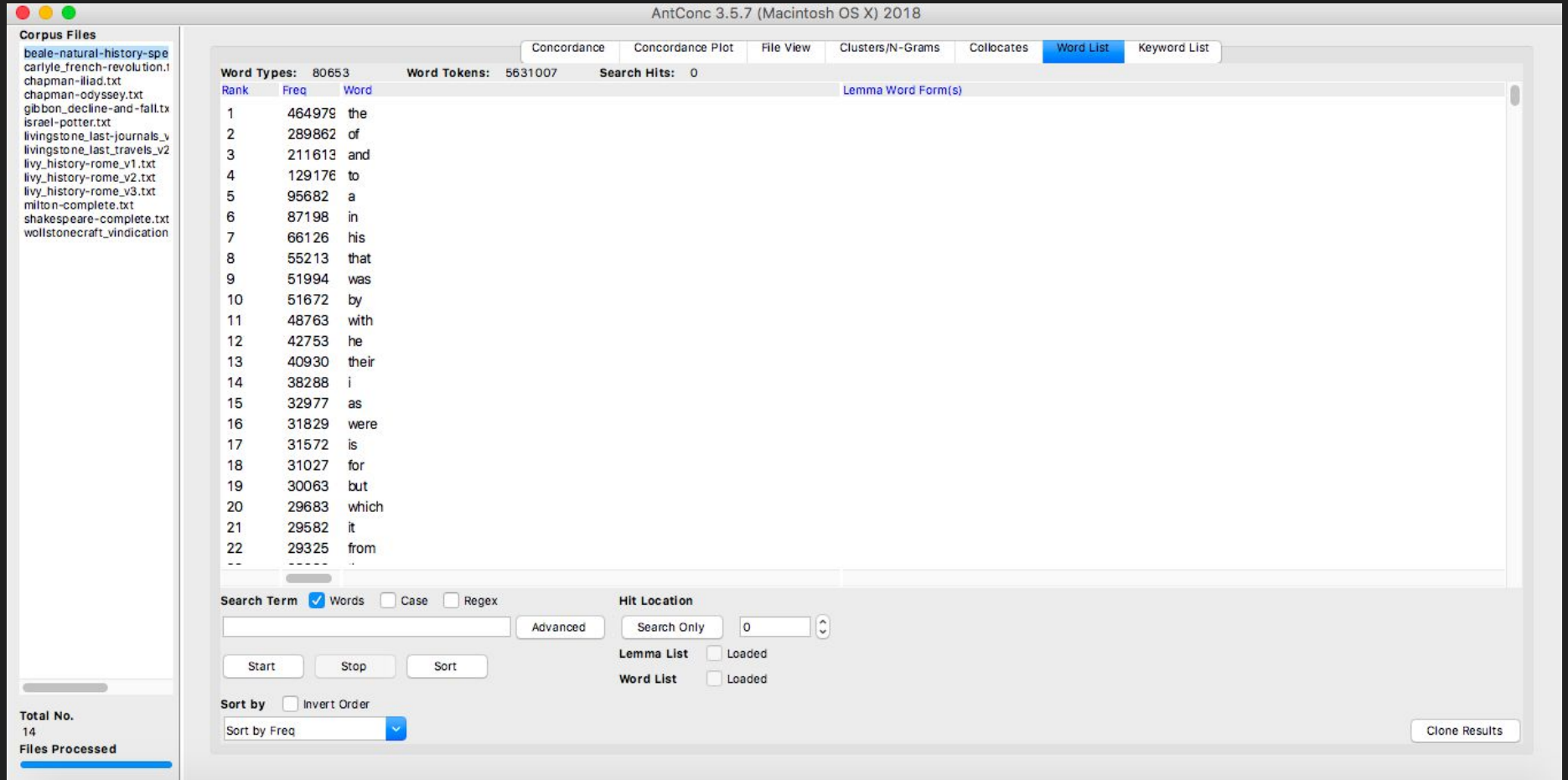
Choose a corpus folder with texts



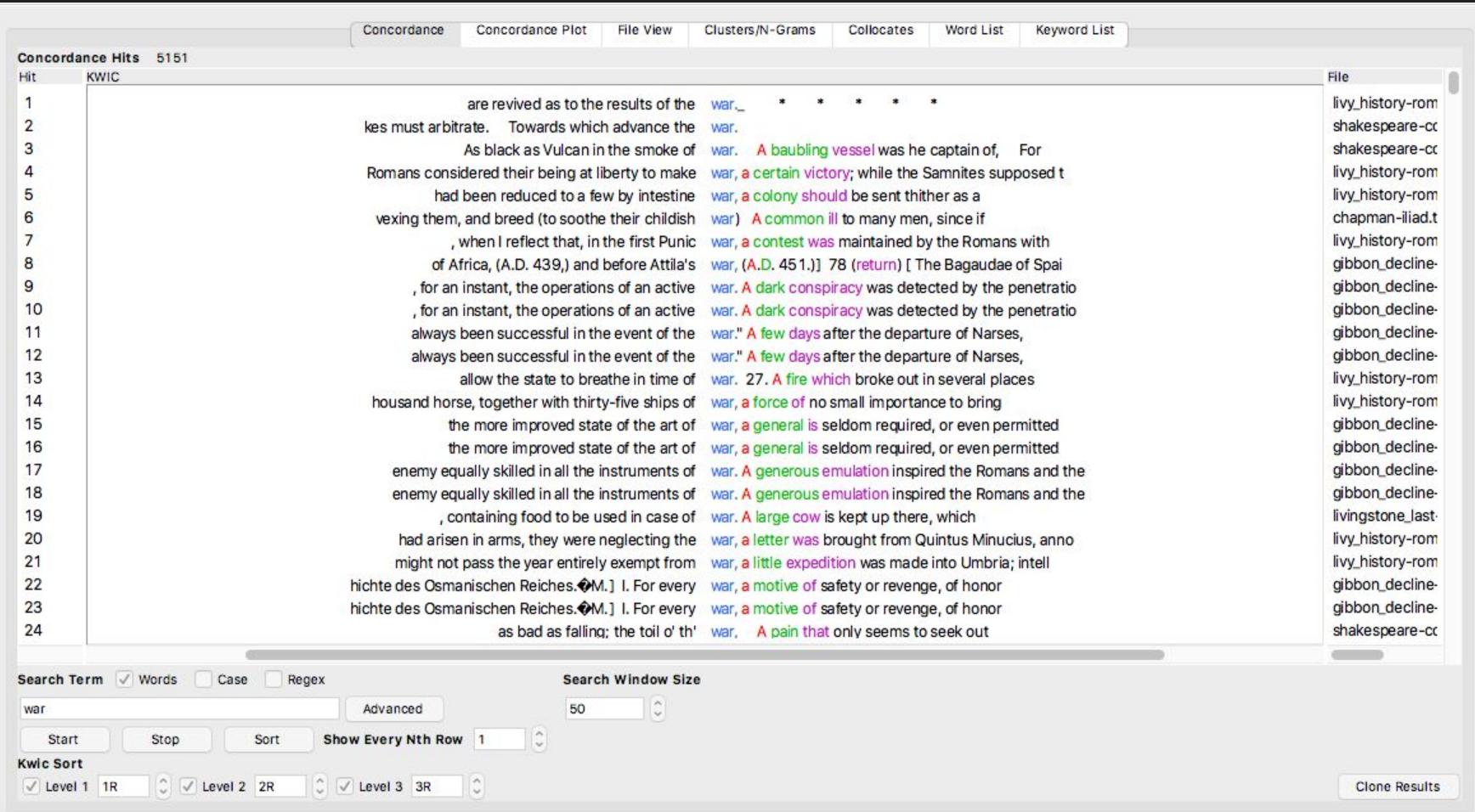
After texts have loaded, click the Word List tab (top right), then click Start (lower left) to generate a word list



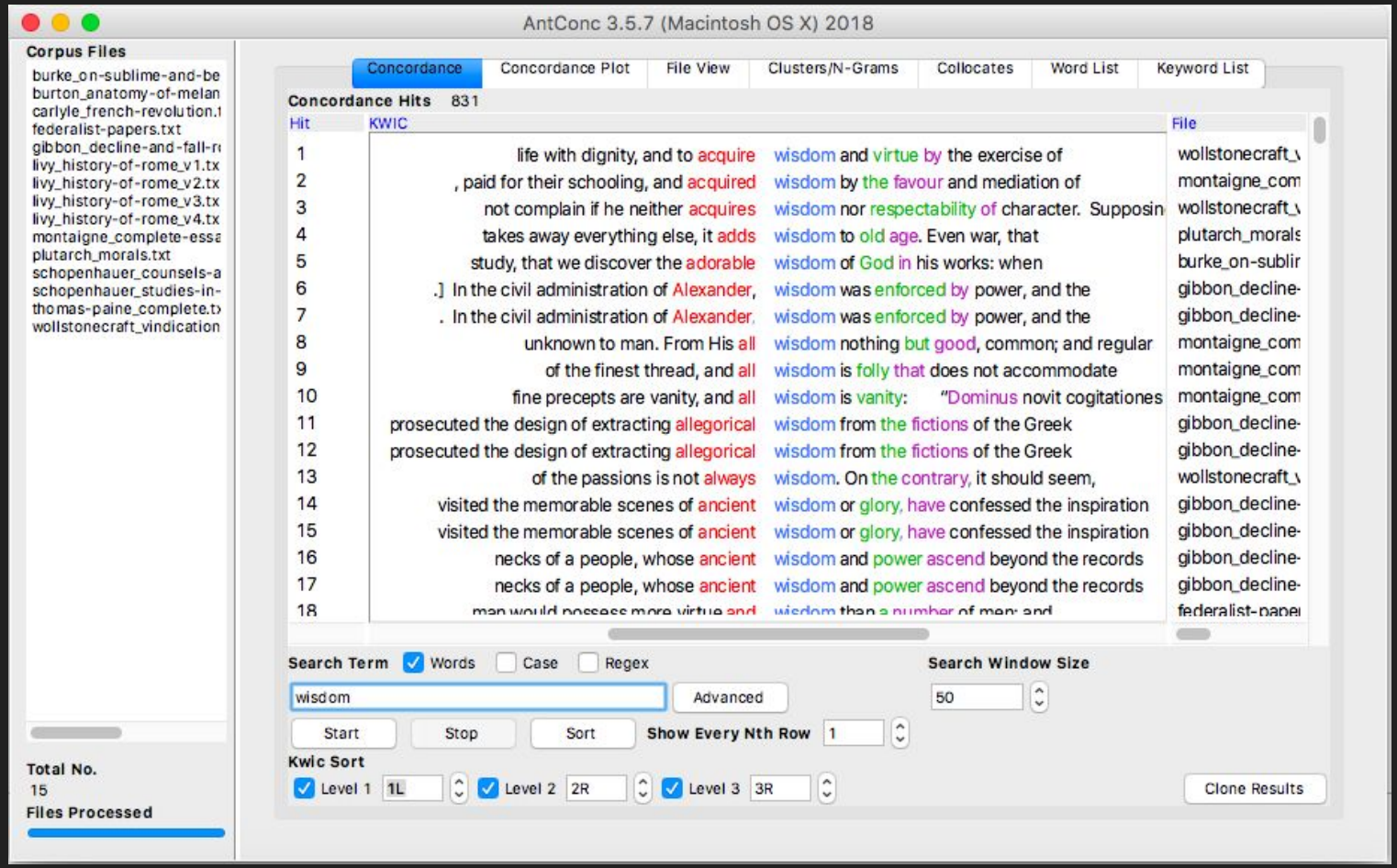
This is what you should get



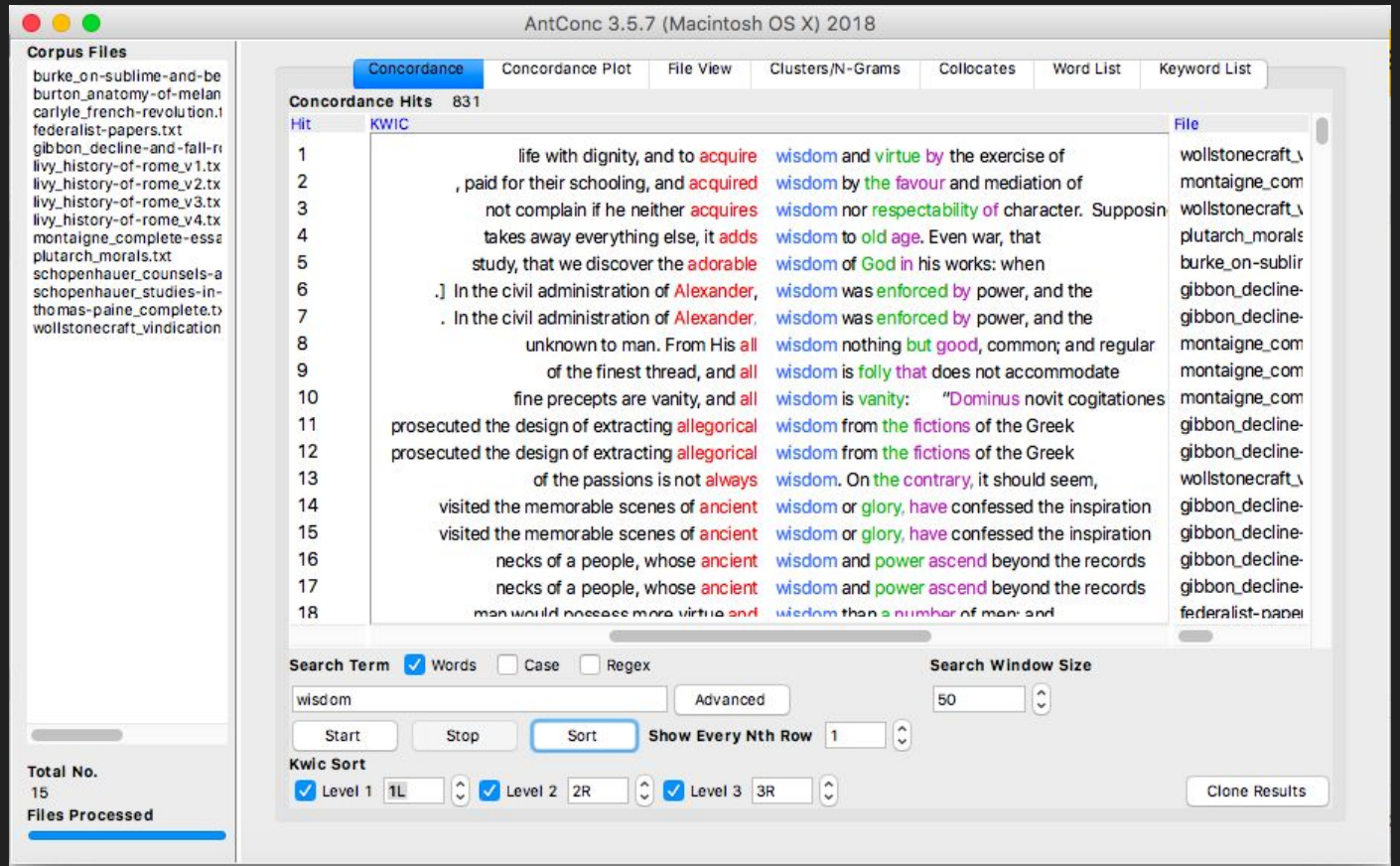
Now click on the Concordances tab and enter a search term



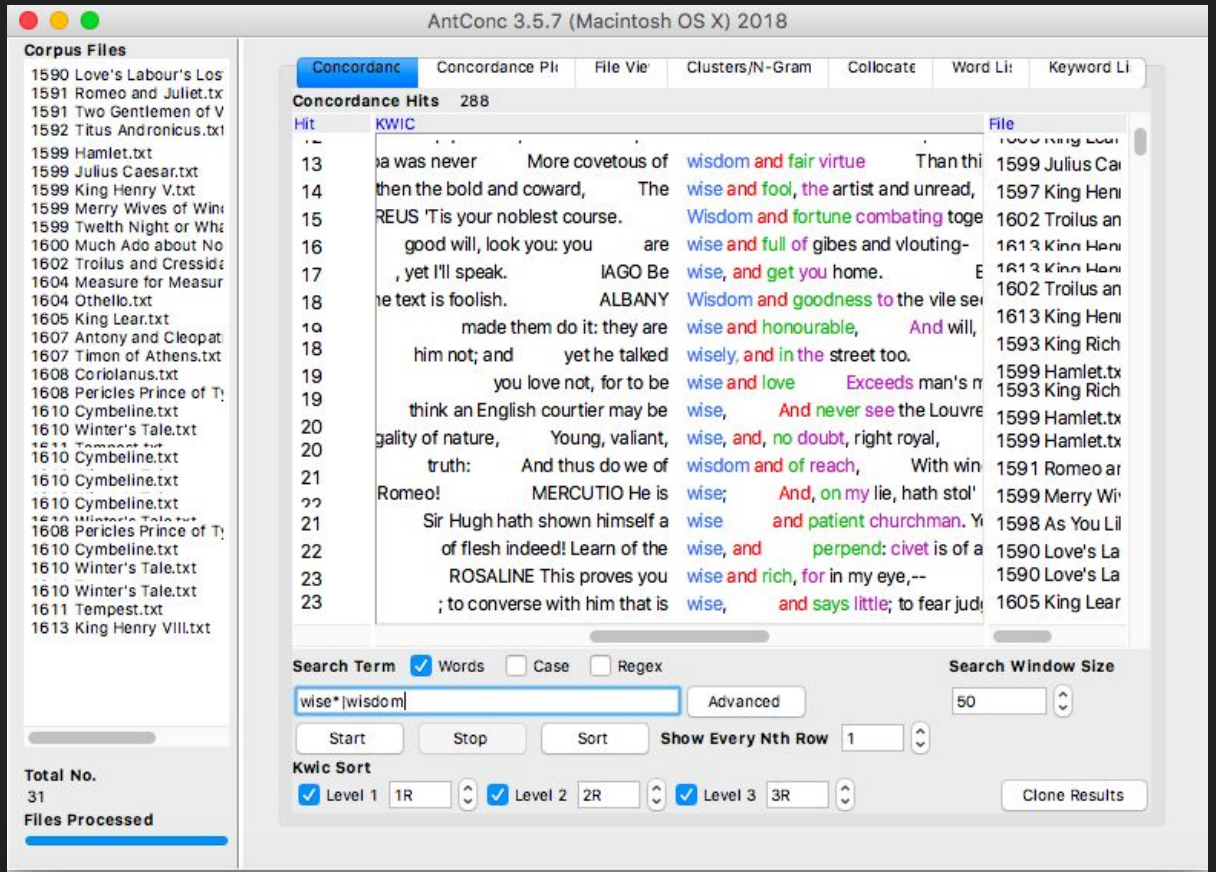
Adjust the concordance results from right to left



Use the “Sort” button to arrange alphabetically



Use regular expressions to do flexible searches: wildcard (*), ? and pipe (|)



More on wildcards

Wildcards are used for matching patterns. Technically they are characters that can be used as a substitute for any class of characters in a search, which increases the flexibility and efficiency of searches.

For a full list of available wildcard operators and what they mean, go to Global Settings > Wildcard Settings.

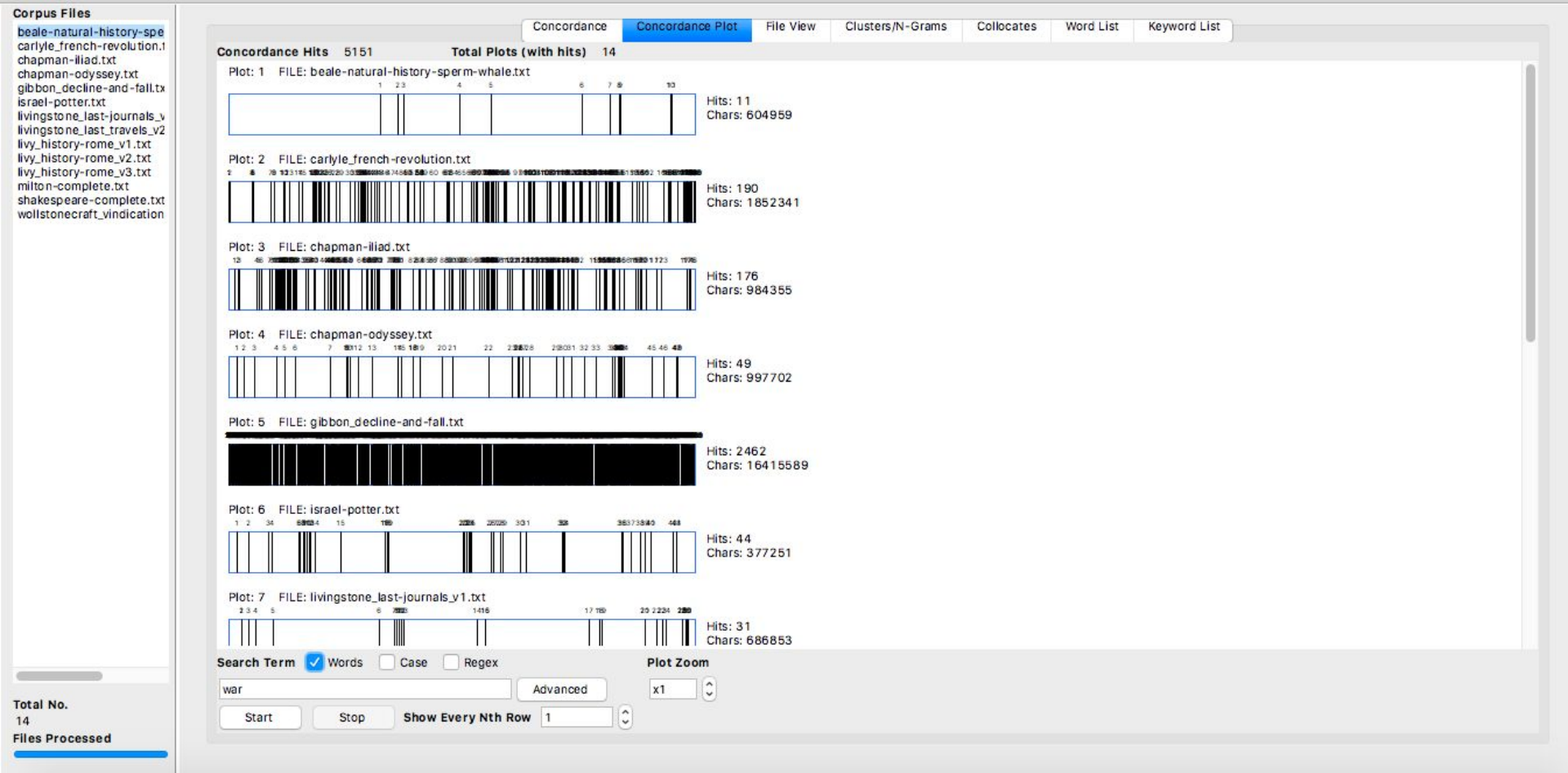
The ? operator is “less greedy” than the * operator:

wom?n – both women and woman

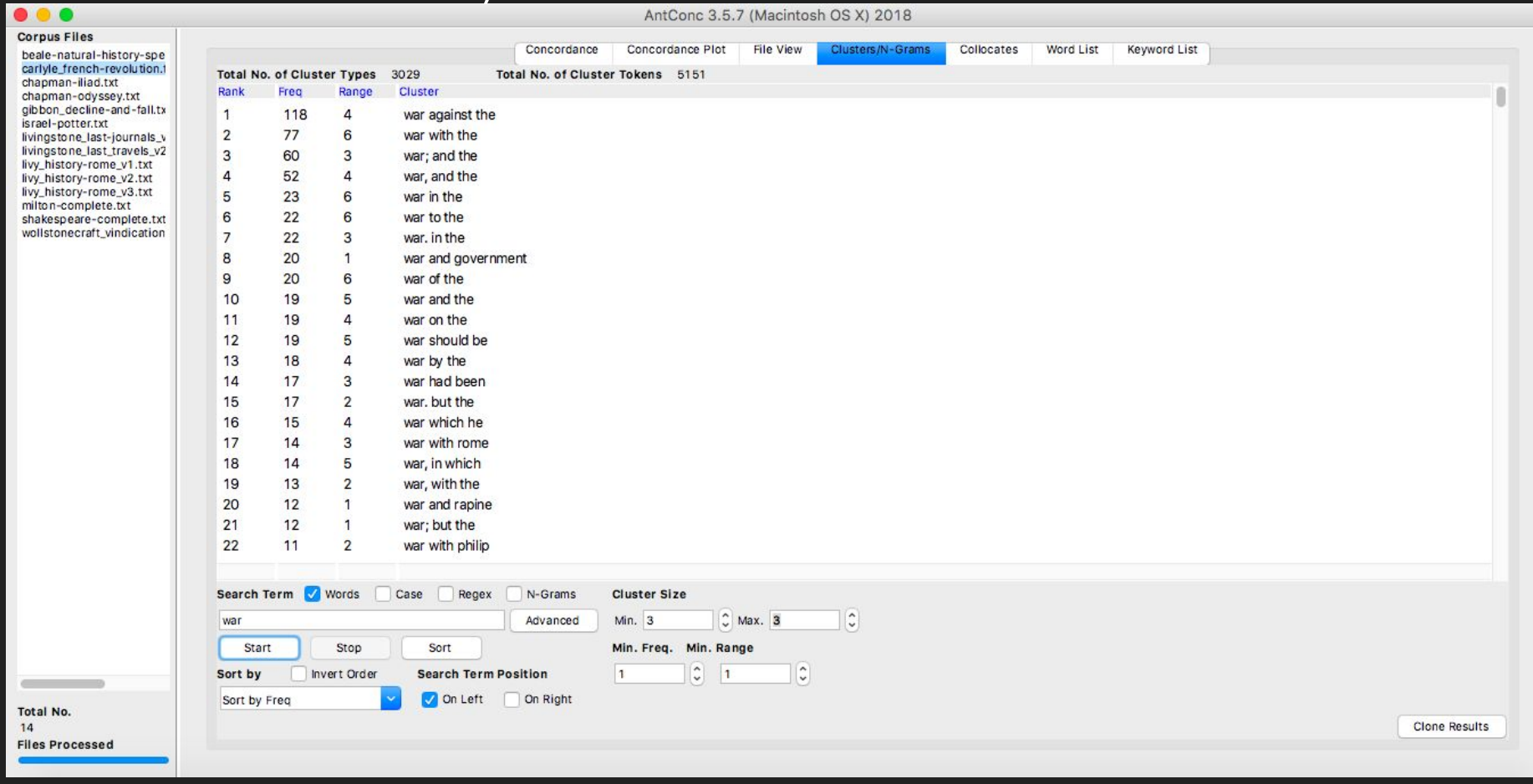
m?n – man and men, but also min

m*n is more flexible: you’ll get mean, melon, etc.

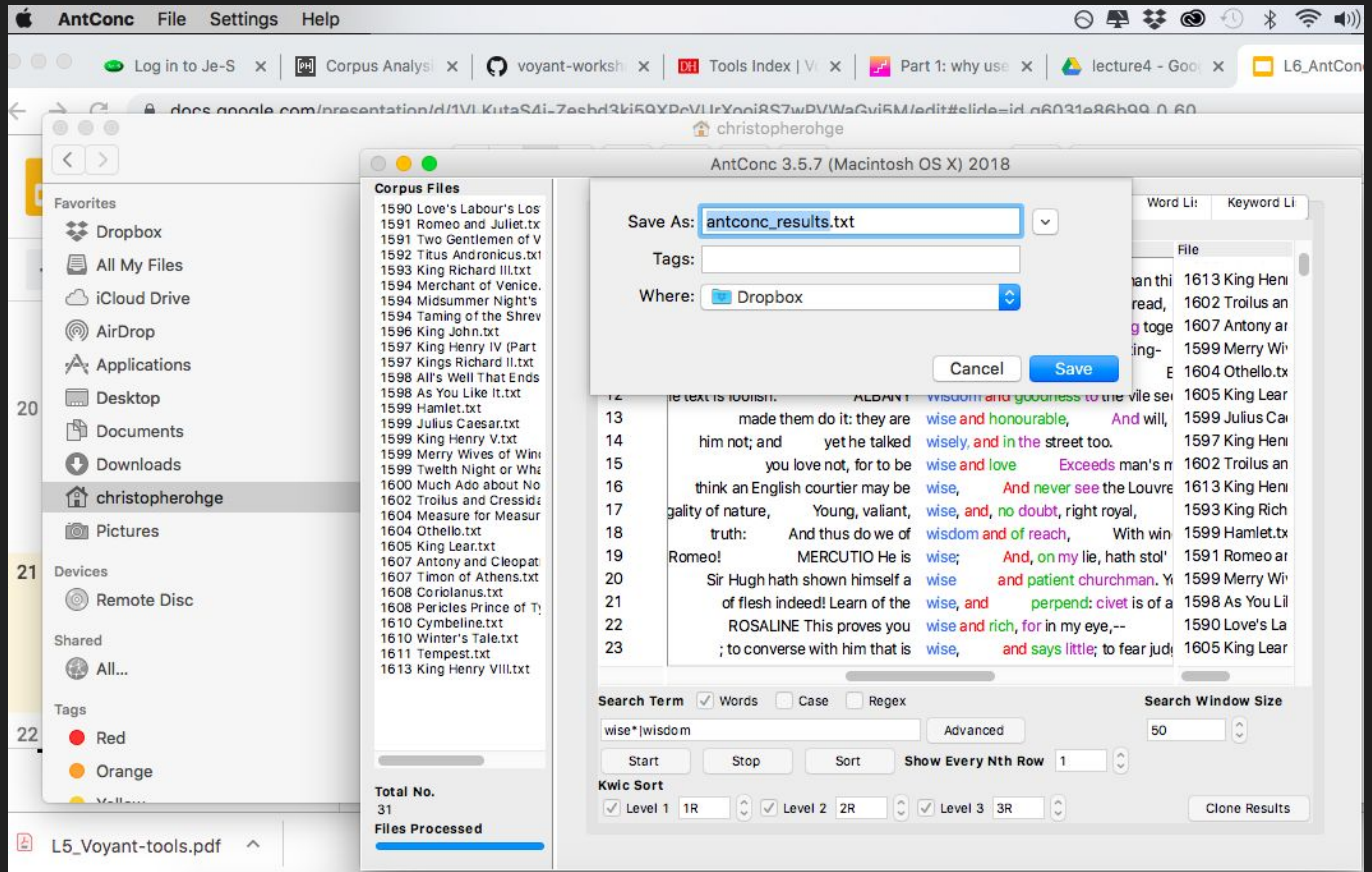
The Concordance Plot tab visualises the search results in each file



Ngram search finds groups of co-occurring words (adjust the “cluster size” to see more context)



Save the output as a txt file



```
1     grace has given a precedent of    wisdom          Above all princes, in
committing freely    1613 King Henry VIII.txt    30  3
2     in nature,          As by your safety,   wisdom, all things
else,          You mainly were  1599 Hamlet.txt 13  13
3     may believe: censure me in your   wisdom, and          awake your
senses, that you   1599 Julius Caesar.txt  14  6
4     Benedick.                    LEONATO O, my lord,  wisdom and blood
combating in so tender            1600 Much Ado about Nothing.txt 18  4
5     , in her sex, her years, profession,              Wisdom and constancy,
hath amazed me more              1598 All's Well That Ends Well.txt  11  3
6     , quarter'd, hath but one part     wisdom          And ever three parts
coward, I     1599 Hamlet.txt 13  11
7     : Saba was never          More covetous of    wisdom and fair
virtue          Than this pure     1613 King Henry VIII.txt    30  12
8     then the bold and coward,          The   wise and fool, the artist and
unread,              1602 Troilus and Cressida.txt   19  1
9          THYREUS 'Tis your noblest course.           Wisdom and fortune
combating together,          If that    1607 Antony and Cleopatra.txt   23  3
10    good will, look you: you          are    wise and full of gibes and
vlouting-    1599 Merry Wives of Windsor.txt 16  9
```

Compare corpora

First, click File > Clear all Tools and Files. Then:

Settings > Tool preferences > Keyword List

Under ‘Reference Corpus’, click “Use raw files”

Add Directory > select the folder containing the text files

Click Load; when it is finished, click Apply.

Click on the Keyword List tab, and click Start. (AntConc will warn that it needs to jump to the Word List—that is fine.) (Note on *Keyness*: this is the frequency of a word in the text when compared with its frequency in a reference corpus, “such that the statistical probability as computed by an appropriate procedure is smaller than or equal to a p value specified by the user.” – taken from here.)

Exercises

1. Load the poetry corpus and sort the hapax legomena alphabetically.
2. Now sort the poetry results to find the most common function words in the poetry corpus.
3. Load the Shakespeare corpus into the program. You should still have the poetry files. How has the function words changed? Perform a basic word search.
4. In the Shakespeare corpus, generate collocates for m?n and wom?n. Now sort them by frequency to 1L.
5. In Whittier's anti-slavery poems, find the names of the poems that talk about female slaves.
6. What are the most commonly negated words (that is, words preceded by "no", "not" and "never") in the c19-20 prose corpus?

BONUS: pos-tagged corpora

Download TagAnt at <http://www.laurenceanthony.net/software/tagant/>.

Open the same corpus directory in the same way as you would with AntConc.

Click start and the program will generated new text files that are annotated with part-of-speech tags.

(We have made the Whittier and *Moby-Dick* pos-tagged files available on the github repository at https://github.com/cmohge1/riga-text-analysis.)

Now run a tagged file through AntConc and see how your results differ (hint: adjust the Kwic sort feature to 2L in order to sort by POS tag).

What is the most common verb form in the King James Bible?